Using NCBI RefSeq Functional Elements to interpret molecular effects of genetic variation in human coronavirus entry gene regulatory regions

Catherine M. Farrell, Terence D. Murphy, and the RefSeq Curation and Development Team

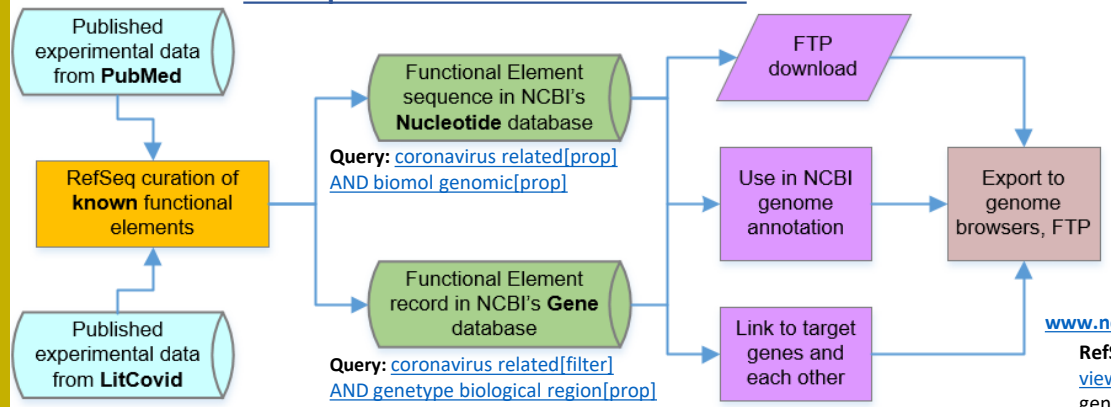
National Center for Biotechnology Information (NCBI), National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

ASHG VIRTUAL MEETING 2020 OCTOBER 27-30

Abstract

The human non-coding genome contains many regulatory elements linked to disease-associated genetic variation but identifying those regions and assessing molecular effects of variation within them is difficult. Consequently, variant interpretation has heavily focused on variants located within genes, particularly within coding regions. The COVID-19 pandemic has drawn attention to the human genes associated with SARS-CoV-2 infection, including genes directly associated with SARS-CoV-2 entry, e.g. *ACE2* encoding the spike protein receptor or *TMPRSS2* encoding a protease that facilitates viral entry, and genes known to be associated with other coronavirus infections. Because these genes and their regulatory regions are of interest for COVID-19 prognosis, NCBI has prioritized annotation of their experimentally validated regulatory regions in the RefSeq Functional Elements dataset (www.ncbi.nlm.nih.gov/refseq/functionalelements/), as recently announced. We provide richly annotated RefSeq records and descriptive Gene database records for these elements (query: coronavirus related[filter] AND genetype biological region[prop]), with annotated features including promoters, enhancers, protein binding sites, among others. Each annotated feature contains concise functional and experimental evidence with links to publications. The dataset is available for FTP download and can be viewed in multiple genome browsers via a track hub (RefSeqFE Hub) that includes regulatory element-to-target gene linkages in bigInteract format. In addition to providing insights into regulatory mechanisms for these genes, these annotations can be used to identify and interpret regulatory variants, either by graphical display of views with custom variation data or by computational means. We overlapped this focused set of regulatory elements with variants from the dbSNP, dbVar, ClinVar and GWAS Catalog databases, and found thousands of variants of **potential** interest to COVID-19 research, including *ABO* intronic variants recently shown to be associated with COVID-19 severity by GWAS (PMID:32558485). In many cases the functional metadata from our feature annotation could inform the molecular effects of overlapping variants, including at transcription factor binding sites and other critical regulatory regions. A similar approach can also be applied to variants associated with other human diseases, whereby RefSeq Functional Elements link regulatory variants to the experimental literature to uncover molecular effects of genetic variation.

RefSeq Functional Elements workflow

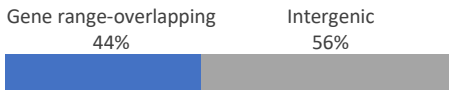


Data access details at: www.ncbi.nlm.nih.gov/refseq/functionalelements/#Data_Access

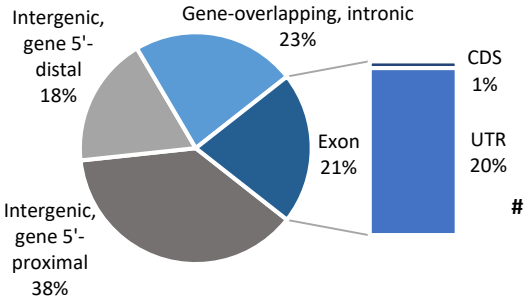
Coronavirus-related data and statistics

Target gene	Regulatory features
ABO	60
ACE2	23
ANPEP	24
CD209	28
CLEC4G	5
CLEC4M	3
CTSL	42
DPP4	5
TMPRSS2	46

Feature gene-relative locations

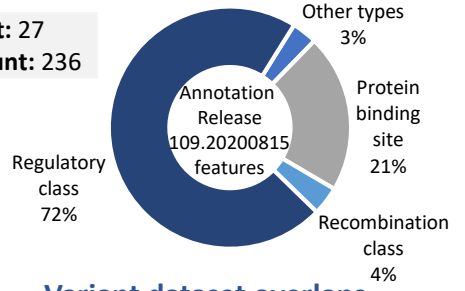


Detailed locations:



Locus count: 27
Feature count: 236

Feature distribution

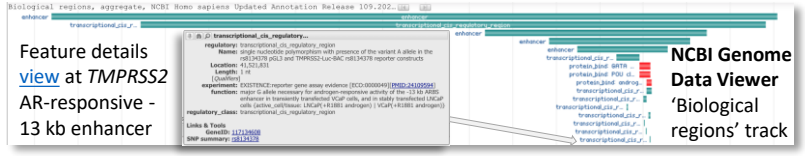


Examples:

1. *LOC112679198*, *ABO* +5.8 intron 1 enhancer, which overlaps 15 severe COVID-19 risk variants from the PMID:32558485 GWAS (95% credible subset); see variant overlap graph
2. *LOC117134604*, *TMPRSS2* promoter region, which includes a G-quadruplex that binds anti-viral small molecule drugs (PMID:32376987)
3. *LOC117134608*, *TMPRSS2* androgen receptor-responsive -13 kb enhancer, which has an AR binding site SNP (rs8134378) that affects enhancer activity (PMID:24109594); shown below

Data graphical displays

www.ncbi.nlm.nih.gov/refseq/functionalelements/#Access_via_NCBI_Graphical_Displays

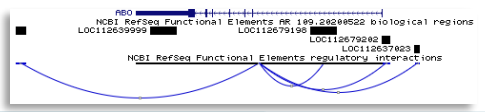


Feature details view at *TMPRSS2* AR-responsive -13 kb enhancer

NCBI Genome Data Viewer 'Biological regions' track

www.ncbi.nlm.nih.gov/refseq/functionalelements/#Access_via_Track_Hub

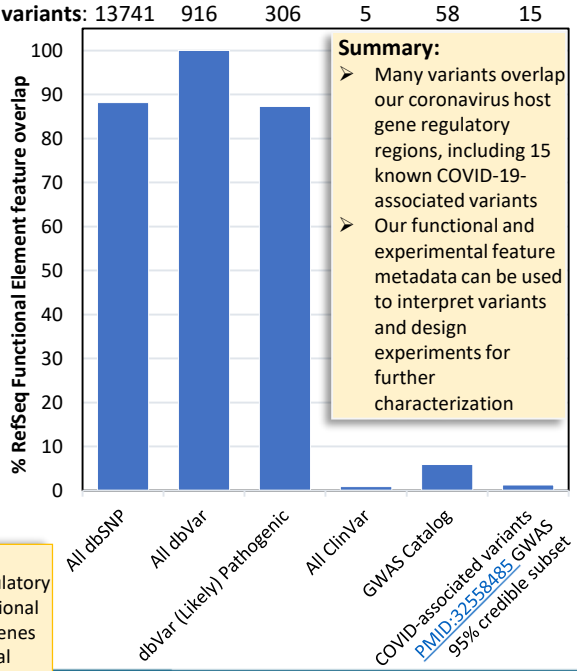
RefSeqFE Hub view of *ABO* gene regulatory interactions



Coming soon!

- Annotation of regulatory elements for additional coronavirus host genes
- A RefSeq Functional Elements publication

Variant dataset overlaps



Summary:

- Many variants overlap our coronavirus host gene regulatory regions, including 15 known COVID-19-associated variants
- Our functional and experimental feature metadata can be used to interpret variants and design experiments for further characterization

Find many more details on our webpage: www.ncbi.nlm.nih.gov/refseq/functionalelements/

Find news about our dataset in the NCBI Insights blog: ncbiinsights.ncbi.nlm.nih.gov/tag/refseq-functional-elements/

Have questions or suggestions? Email: farrelca@ncbi.nlm.nih.gov

Nothing to Disclose

This work was supported by the Intramural Research Program of the NIH, National Library of Medicine

ASHG American Society of Human Genetics